Clim. Past Discuss., 2, S156–S160, 2006
www.clim-past-discuss.net/2/S156/2006/

Climate
of the Past
Discussions

Interactive comment on “On the verification of climate reconstructions” *by* G. Bürger and U. Cubasch

G. Bürger and U. Cubasch

Received and published: 7 July 2006

Comment on "Interactive comment on ’On the verification of climate reconstructions’" by G. Bürger and U. Cubasch

This comment represents a quite comprehensive overall criticism of a number of studies, including published ones, undertaken by several research groups which have partly been responded to elsewhere. We will try to isolate those points that directly affect our paper, and a few others (marked by "(*)") that we find worth discussing.

1."decade old work". - That decade old work - the MBH98 reconstruction - originated the mentioned NAS panel, to address, among other things, the very issues of our paper. And according to Rutherford et al., 2005, Table 2, for that decade old work an even better millennial skill is estimated than for the newer RegEM version (RE of 0.51

vs. 0.40), if calibration period discrepancies between both sources can be ignored (1901-71 vs. 1902-80). Moreover, we explicitely included this newer version into our framework (flavor 0130), resulting in an RE score of 0.30 using the classical calibration period.

2.(*) The detrending issue of von Storch et al., 2004 was first noted by Bürger et al., 2006 (B06, and as early as January 2005 rejected as a Science comment). The conclusions of that study and of Bürger and Cubasch, 2005 (B05) are not invalidated but instead supported by this issue, as it increases the overall sensitivity of the method to data processing details.

3."RegEM". - The rev. claims that RegEM is misrepresented in B05/6 as it contains an error covariance estimation. This proves a thorough misunderstanding of B05/6 which deal with the error in model >coefficients<. The current paper has a different topic, so it remains unclear why it misrepresents RegEM. Moreover, the rev. fails to react to our RegEM issue in the supplement, that is, why RegEM should be applied at all, instead of the classical EM procedure (the intermediate regression step is not ill-posed if 22 proxies are used).

4."legitimate flavors". - (GLB) From an algorithmic viewpoint there is no difference between using all temperature grid points or some EOF projections/averages thereof as a target variable. It is simply not convincing if the final result depends on that specific choice. Note that MBH98 describe the EOF reduction as the second of three of its "fundamental assumptions". (MDL) Likewise, although ridge regression, as Schneider, 2001 puts it: "might still offer advantages" over using truncated total least squares, it is not convincing if the result changes when replacing one by the other. We have included the latter simply because it was one of the options in the original Schneider algorithm (see http://www.gps.caltech.edu/˜tapio/imputation/regem.m). (RSC) The rev. believes that rescaling "makes absolutely no sense" in the context of RegEM, ignoring that RegEM, like in any other regression based model, leads to variance attenuation (cf. Schneider, 2001, p. 868). Therefore, in a calibration/verification exercise rescaling is as reasonable to do with RegEM as it possibly was for MBH98, and suboptimality also applies to both of them.

The rev. repeatedly bases his/her arguments on the uselessness of the resulting flavors. But for all of these options there are arguments pro and con, and none of them is a priori valid. On the other hand, if one feels uncomfortable with a specific flavor one might simply ignore it. As long as there are some "legitimate" flavors left (such as 1011 or 0130, see below) our main conclusions apply.

5.(*) The rev. has a misconception of principal component regression, assuming the EOF transformation for the target (temperature) instead of the predictors (proxies).

6.The rev. has not correctly updated his/her GRL review of our original version. There we incorrectly reported a 57% verification RE for the NH mean (it was the global mean), and corrected this. Apart from this, nothing else of the GRL review was helpful.

7.Our emulations utilize two unique datasets, the 22 proxies of the AD 1400 network and the 219 temperature gridpoints selected by MBH98 for the verification of NH mean. That we used this reduced set of grid points even for calibration was one of the main criticisms by the other GRL reviewer. It is undisputed that the enlarged set yields a higher RE score, but since that set is not observed in the earlier period, the corresponding model is not fully verifiable in our bootstrapping framework guided by stationarity. On the other hand, the reduced set was assumed by MBH98 to represent the NH mean in the verification part. Should this not be the case for the calibration would certainly not support the robustness of the method.

8."Rutherford et al., 2005". - Rutherford et al., 2005 have used, compared to MBH98, a) different temperature data (NH only, newer CRU version), b) different calibration periods than MBH98 (1901-1971 and 1856-1928), and c) an infilling via RegEM of the incomplete temperature data set prior to the entire analysis. If the method is robust these differences have no effect on the verification, with one possible exception: While c) is not a problem per se it introduces a verification bias. Because the infilling uses covariance information from the entire period (as the article suggests), any calibration uses information from the validation set which is thus no longer independent. For this setting they report a verification RE of 40% (66% for the "early" calibration, with even 71% for the "hybrid-20" variant), as opposed to our emulated RE of 30%. Hence, this discrepancy is indicative either of the mentioned bias, or of the enlarged number of grid points, or of the insignificance of the RE measure, or of all of them together, but in any case of the non-robustness (uncertainty) of the verification measure.

That the RE is increased when using the late verification does not contradict our claim, as the rev. incorrectly states. Strongly autocorrelated series inflate RE and its level of significance, especially a trend over the entire period.

Regarding the hybrid-20 variant, Rutherford et al., 2005 state: "As described below, cross-validation experiments motivate the choice f = 0.05 cpy (20-yr period) for the split frequency in almost all cases." - This sentence illuminates how a posteriori selections penetrate a model definition. If there is no extra (independent, a priori) motivation for that choice, the chosen model is no longer independently verifiable. This is especially critical if the above cross-validation is sensitive to the parameter in question, and is of course dramatically aggravated with the number of such choices.

9."significance". - Following the rev.'s opinion about the RE significance level, a nonsense predictor such as the number of available grid points has a significantly nonzero predictive RE skill. (I assume here that we weren't so lucky to pick the right one out of, say, 100 - it was just one try). It appears that we (along with McIntyre and McKitrick, 2005) and the rev. (along with Mann et al., 2005 and Huybers et al., 2005) use the phrase "RE significance" in a totally different sense. In the latter case, zero skill is defined by the skill of a purely stochastic red noise predictor. In the former case, zero skill is defined after that stochastic predictor is sent through a regression, so that amplitude and sign are adjustable. It is not difficult to judge which of the two versions is more adequate for our nonsense predictor (the number of available grid points as a $>$regressor$<$).

10."cross validation". - We have emphasized that calibration and validation must not reflect mere sample properties. Rutherford et al., 2005 go one step into the right direction by swapping calibration and validation sets (but then it is unclear how the special calibration period 1856-1928 is motivated). To rigorously implement the above condition the selection process must be fully randomized.

11."Mann et al., 2005". - The rev. tries to disprove our conclusions citing Mann et al., 2005. That study nicely demonstrates a successful reconstruction of a simulated climate history from sufficiently many pseudoproxies (104, representing the AD 1780 network), which obviously cannot contain any nonsense predictors and which has never been questioned by us. Our focus was on the AD 1400 network with 22 real proxies.

Interactive comment on Clim. Past Discuss., 2, 357, 2006.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper